Dublin Business School
excellence through learning

QQI

Data and Web Mining

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS

B8IT108_TMD3_1718

Module Code: **B8IT108**

Module Description: **Data and Web Mining**

Examiner: **Terri Hoare**

Internal Moderator: **Darren Redmond**

External Examiner: **Dr Brett Becker**

Date: 24 September 2018
Time: 10:00 -12:00

INSTRUCTIONS TO CANDIDATES

Time allowed is 2 hours
Question 1 is mandatory (30 marks)
Answer 2 of the remaining 3 Questions (35 marks each)

All answers should reference literature and case studies as appropriate. Use of scientific calculators only is permitted.

Question 1 : MANDATORY **(30 Marks)**

a) The AllElectronics Customer Database Table below includes attributes age, income, student, credit_rating, and buys_computer (yes/no).

Information Gain is one of the methods used by a supervised decision tree algorithm. It can be applied to determine which attributes to use for a tree-split so that uncertainty in the prediction buys_computer (yes/no) is minimised.

Calculate Information Gain for the Age attribute using the data in the AllElectronics Database Table.

Hint: apply the formula for Information Gain.

$$Information\ Gain(A) = Info(D) - Info_A(D)$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Electronics Customer DB - Training Tuples - Label Buys_computer

RID	**Age**	**Income**	**Student**	**Credit_rating**	**Buys_computer**
1	youth	high	no	Fair	**NO**
2	youth	high	no	excellent	**NO**
3	mid-aged	high	no	Fair	**YES**
4	senior	medium	no	Fair	**YES**
5	senior	low	yes	Fair	**YES**
6	senior	low	yes	excellent	**YES**
7	middle-aged	low	yes	excellent	**YES**
8	youth	medium	no	Fair	**NO**
9	youth	low	yes	Fair	**YES**
10	senior	medium	yes	Fair	**YES**
11	youth	medium	yes	excellent	**YES**
12	middle-aged	medium	no	excellent	**YES**
13	middle-aged	high	yes	Fair	**YES**
14	senior	medium	no	excellent	**NO**

[10]

b) A model has been built to predict the risk of counterparty credit default by training a Support Vector Machine (SVM) model on credit default data. The confusion matrix below was generated on test data for the trained model. Interpret the reported accuracy score and calculate and interpret the related performance measures of recall, precision, specificity, and sensitivity.

accuracy: 95.64% +/- 3.25% (mikro: 95.64%)

	True No	True Yes
Predicted No	286	10
Predicted Yes	7	87

[10]

c) Appraise the Davies-Bouldin Index as a performance measure for clustering algorithms.

[5]

d) Appraise the role of feature selection in the data preparation phase of the CRISP-DM data mining framework.

[5]

ANSWER (2 OUT OF 3) QUESTIONS BELOW

Question 2 **(35 Marks)**

a) Select and critique an appropriate algorithm for building a model using supervised learning. Your answer should include a high-level description of how the algorithm works, input and output limitations, pros, cons, and types of suitable applications.

[10]

b) Supervised learning algorithms can be grouped as either ‘bucketing’ or function fitting algorithms. Motivate the reasons that logistic regression can be considered both a ‘bucketing’ and a function fitting classifier.

[6]

c) Describe and critique cross validation as a method for evaluating the training performance of a model.

[8]

d) What is meant by ‘overfitting’ a model?

[3]

e) Most of the data mining models deployed in production applications are ensemble models. In an executive board analogy, having a board with diverse and independent members makes statistical sense. Propose four methods of achieving diversity in the base models making up an ensemble.

[8]

Question 3 **(35 Marks)**

a) Propose a data mining taxonomy for unsupervised learning problems. Include examples of data mining algorithms.

[10]

b) Select and critique an appropriate algorithm used for clustering data. Your answer should include a high-level description of how the algorithm works, input and output limitations, pros, cons, and types of suitable applications.

[10]

c) Appraise the Euclidean and Jaccard methods for calculating ‘nearness’ of n-dimensional data points in clustering algorithms.

[6]

d) Appraise the role of the support and confidence thresholds in identifying interesting rules in Association Analysis.

[4]

e) Propose an algorithm for identifying frequent itemsets in Recommender Engines. Give a brief description of how the algorithm works.

[5]

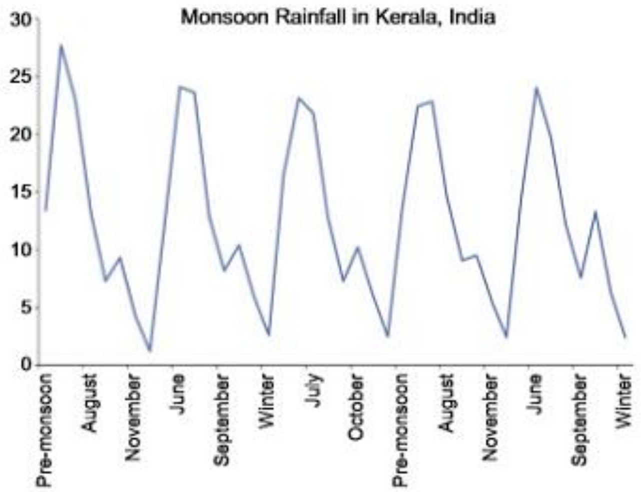
Question 4 **(35 Marks)**

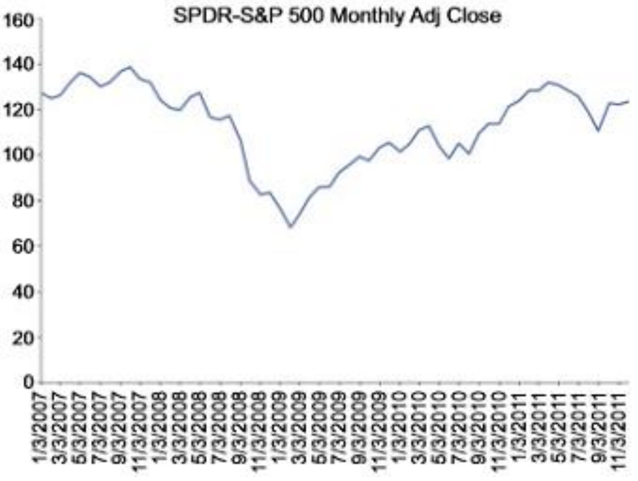
a) PageRank is an algorithm used by Google Search to rank websites. Critique the measure of network centrality that it uses and contrast with at least two other measures of 'centrality' or power in a network graph.

[15]

b) The figures below show two different time series: annual monsoon precipitation in Southwest India averaged over a five-year period and the adjusted month-end closing prices of the SPDR S&P 500 (SPY) Index over another five-year period. For each time series, propose either a data or model driven forecasting approach motivating the selection of the approach in each case.

[8]





c) Discuss the role of TF-IDF in text mining.

[6]

d) Propose a supervised learning approach for performing sentiment analysis of blogs.

[6]

END OF EXAMINATION